Informatica e diritto, XXXIX annata, Vol. XXII, 2013, n. 1, pp. 263-280

Text and Social Network Analysis as Investigative Tools: A Case Study

NICOLA LETTIERI, DELFINA MALANDRINO, RAFFAELE SPINELLI*

1. INTRODUCTION

This paper explores the intersections between the law and the computational social science (CSS) paradigm by focusing, in particular, on text and social network analysis. We will present ongoing research about the applications of computational methods in the analysis of structural and functional features of criminal organizations. Inspired by a sociological study using network analysis techniques to compare the characteristics of two criminal organizations belonging to Sicily's *mafia* and Campania's *camorra*, the research aims at studying tools combining information extraction, network analysis and visualization methods to support investigation and the fight against criminal organizations. After a brief introduction to social network analysis (SNA) and its applications in the legal field, the paper offers an overview of the results so far achieved from a technical and methodological point of view sketching future developments that appear to be challenging both for criminology and legal informatics.

* N. Lettieri, researcher at ISFOL, Rome, is adjunct professor of Legal informatics at the University of Sannio, Benevento (Italy) and of Computational social sciences at the Department of Computer science of the University of Salerno (Italy); D. Malandrino and R. Spinelli are, respectively, assistant professor and PhD student of Computer science at the University of Salerno (Italy). The Authors would like to thank the following people for their useful contribution, suggestions and comments: Attilio Scaglione, University of Palermo; Carlo Rinaldi, deputy prosecutor, Public Prosecutor's Office of Sala Consilina (Salerno); Luigi Landolfi, deputy prosecutor of the Antimafia District Department (DDA) of Naples. The paper is the result of joint collaboration: Sections 1-4 and 6 were written by N. Lettieri; D. Malandrino and R. Spinelli, responsible for the technological aspects of the project, are the Authors of Section 5.

2. Social Network Analysis

Social network analysis is a theoretical and methodological approach to the study of social phenomena[1] whose origins can be traced back to the 1930s. Ideally stemming from the intuitions of the sociologist Georg Simmell[2] and of the psychosociologist Jacob Levi Moreno[3], SNA aims at analyzing and understanding forms of social life as distinct from their content: instead of focusing on individuals and their attributes (gender, age, instruction, economic status, opinions, etc.), or on macro social structures, it centers on the *relations* between individuals, groups, or social institutions.

In the SNA perspective, to study society is to study social actors within the network of relations in which they are emerged, seeking explanations

[1] The birth of SNA is connected with the growth, over the 20th century, of different scientific communities sharing the same basic idea: to understand human and social phenomena focusing more on the role played by social structures than on individuals and their attributes. Among the most relevant schools of thought, we can mention the School of Sociometry lead by Jacob Moreno, the Harvard Sociological School of Lloyd Warner and the anthropological research group at the University of Manchester (Max Gluckman, John Barnes). The bibliography on network analysis is extremely wide. Among others see: B. Wellmann, S.D. Berkowitz (eds.), *Social Structures: A Network Approach*, Cambridge, Cambridge University Press, 1988; S. Wassermann, K. Faust, *Social Network Analysis, Methods and Applications*, New York, Cambridge University Press, 1997; D. Knoke, S. Yang, *Social Network Analysis*, Thousand Oaks, Sage, 2008. Interesting insights about the role of networks in social dynamics are offered in: A.-L. Barabási, *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life*, New York, Plume Books, 2003; N.A. Christakis, J.H. Fowler, *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives*, New York, Little, Brow and Company, 2009.

[2] With his Formal Sociology, Georg Simmel pioneered many concepts of SNA. According to Simmel, relationships that individuals create in their continuous interactions clearly influence the action of the subject. Therefore, the task of sociology is to isolate the forms of social life abstracting them from their concrete content. Simmel wanted to develop a geometry of social life in order to investigate it in a quantitative and operational way: "Geometric abstraction investigates only the spatial forms of bodies, although empirically these forms are given merely as the forms of some material content. Similarly, if society is conceived as interaction among individuals, the description of the forms of this interaction is the task of the science of society in its strictest and most essential sense", see G. Simmel, *The Study of Societal Forms*, in Wolff K.H. (ed.), "The Sociology of George Simmel", Glencoe, The Free Press, 1950, pp. 21-22.

[3] Jacob Levy Moreno is known, among other things, for being the father of Sociometry, the quantitative method for measuring social relationship. In 1937, he began to publish the journal *Sociometry: a Journal of Inter-Personal Relations*. On the sociometric paradigm, see J.L. Moreno, *Sociometry, Experimental Method and the Science of Society. An Approach to a New Political Orientation*, New York, Beacon House, 1951.

for social behaviors in the structure of these networks rather than simply in the individuals. According to this approach, social relationships are conceptualized, represented and studied as graphs, static pictures consisting of nodes (actors) and ties (connections between actors). Once a graph is generated according to a given metric (e.g., mapping friendship, kinship, cultural exchanges, organizational position, etc.), several measures are then used to analyze structural and functional features both of the network and its components. The study of *Centrality*[4], one of the most relevant measures in SNA, is used, for example, to assess issues such as the dominance, subordination, influence or prestige of social actors. Just to give an idea of the meaning and role of SNA metrics, we would like to mention here the main centrality measures: *degree, closeness and betweenness centrality*.

Degree centrality is defined as the number of direct links a node *k* has:

$$C_D(k) = \sum_{i=1}^{n} a(i,k),$$

where *n* is the total number of nodes of the network, and *a(i,k)* is a binary variable indicating whether a link exists between nodes *i* and *k*. *Degree centrality* is used to measure the activity of a particular node: a node with a high degree is likely to be a leader or a "hub" within the group.

Closeness centrality is the sum of the length of the geodesics between a particular node k and all the other nodes in a network and is defined by the expression:

$$C_C(k) = \sum_{i=1}^{n} l(i,k),$$

where *l (i, k)* is the length of the shortest path connecting nodes *i* and *k*. Closeness allows us to measures how far away one node is from other nodes.

Betweenness centrality of a node *k* is the number of shortest paths between all the vertices of the network (geodesics) and passing through it and is defined by the expression:

[4] In network analysis, the centrality of a node within a graph is a measure that allows us to determine the importance of a social actor, that is, how influential a person is within a group. About *centrality* and its measures see L. FREEMAN, *Centrality in Social Networks: Conceptual Clarification*, in "Social Networks", 1979, 1, pp. 215-239.

$$C_B(k) = \sum_{k}^{n} \sum_{j}^{n} g_{ij}(k),$$

where $g_{ij}(k)$ indicates whether the shortest path between two other nodes *i* and *j* passes through the node *k*. The *betweenness* measures the centrality of a node: a node with high *betweenness* is probably a "broker", a good intermediary within communication or market dynamics (flow of goods, opinion spread).

By means of these and other measures[5], SNA is offering new insights for many social sciences, such as sociology, anthropology, economics and has been used effectively in the study of extremely various phenomena[6], finding applications in the physical, biochemical, genetic and computer sciences, while maintaining conventionally the name social in memory of its origin. Even if originally thought of for the analysis of human relationships, SNA metrics and techniques have spread to studying relations between the most diverse entities from documents to computers: the recurrence of the term "SNA" in scientific literature over the last ten years shows an exponential growth in the use of this representation of social phenomena[7].

3. Social Network Analysis in the Legal Field

In its more recent life, SNA has shown interesting potentialities in providing new insights also into legal and socio-legal issues. The application of network analysis techniques to the study of legal documents has become

[5] For a general introduction to SNA measures, metrics and techniques see J. Scott, *Social Network Analysis. A Handbook*, Thousand Oaks, Sage, 2000; M.E.J. Newman, *Networks: An Introduction*, Oxford, Oxford University Press, 2010.

[6] Over time SNA has been applied in the study of extremely various phenomena. See, i.a.: J. Diesner, K.M. Carley, *Using Network Text Analysis to Detect the Organization Structure of Covert Networks*, in "Proceedings of the North American Association for Computational Social and Organizational Science Conference", 2004, http://goo.gl/nwWzM; R. Popping, *Text Analysis for Knowledge Graphs*, in "Quality and quantity", Vol. 41, 2007, pp. 691-709; P.A. Stockowski, *Leisure in Society, a Network Structural Perspective*, London, Mansell Publishing, 1994; B. Erickson, *Culture, Class and Connections*, in "American Journal of Sociology", 1996, n. 102, pp. 217-225; N.A. Christakis, J.H. Fowler, *The Spread of Obesity in a Large Social Network over 32 Years*, in "New England Journal of Medicine" Vol. 357, 2007, pp. 370-379; R.M. Christley, G.L. Pinchbeck, R.G. Bowers *et al.*, *Infection in Social Networks: Using Network Analysis to Identify High-risk Individuals*, in "American Journal of Epidemiology", Vol. 162, 2005, pp. 1024-1031.

[7] See D. Knoke, S. Yang, *Social Network Analysis*, cit., p. 1.

the basis of several studies aiming at capturing the structural and dynamic properties of specific legal phenomena from legislation to case law.

In 2005, Thomas A. Smith from the University of San Diego published on this topic a position paper[8] claiming the possibility of identifying in the relations between legal documents (cases, statutes and other legal authorities, and the citations that link them together) the typical topology of the network structure[9] and, consequently, proposing to apply network analysis measures to law and legal issues. After a brief introduction to the basic concepts of network science, the paper presents the results of citation study based on the analysis of nearly four million American legal precedents, showing how the study of the network made up by cases and their mutual relations could help to shed light on how the legal system evolves. According to Smith, such a kind of analysis allows us not only to empirically measure the degree of integration between different legal systems (e.g., state and federal laws) and its evolution over time but also to analyze the emergence and decline in the authority of precedents over the years. In this perspective, the analysis of cases can be integrated with that of legal authority to verify their mutual relations.

Since 2005, many works focusing on the application of SNA techniques to case law and legislation have been published. As to the analysis of case law, it is worth mentioning, among the others, research from Chandler[10] and Fowler[11], who used network analysis to study relationships between precedents of the Supreme Court of the United States in order to capture information on the relevance of judgments that cannot be detected by simply counting the number of citations[12].

[8] T.A. SMITH, *The Web of Law*, San Diego Legal Studies Research Paper, 2005, n. 06-11, available at SSRN: http://ssrn.com/abstract=642863.

[9] "...the overall topology, or mathematical structure, of the Web of Law closely resembles that of the World Wide Web. Both the World Wide Web and the Web of Law are "directed" networks, have grown organically to a large size, and evince striking features of self-organization", T.A. SMITH, *The Web of Law*, cit., p. 2.

[10] S.J. CHANDLER, *The Network Structure of Supreme Court Jurisprudence*, University of Houston Law Center, No. 2005-W-01, available at SSRN: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=742065.

[11] J.H. FOWLER, T.R. JOHNSON, J.F. SPRIGGS, S. JEON, P.J. WAHLBECK, *Network Analysis and the Law: Measuring the Legal Importance of Supreme Court Precedents*, in "Political Analysis", Vol. 15, 2007, n. 3, pp. 324-346.

[12] See also M.J. BOMMARITO II, D.M. KATZ, J. ZELNER, *Law as a Seamless Web? Comparison of Various Network Representations of the United States Supreme Court Corpus (1791-*

With regard to the analysis of legal texts, we can report a work from Katz and Bommarito[13] on the United States Code of which a representation formalized in terms of a network is proposed in order to measure the direction and the magnitude of the changes suffered by the Code itself over time. The French Environment Code has been the subject of an analysis of the same kind carried out by Boulet and others[14], who considered the network formed by citations in the Code that, as suggested by the analysis conducted, presents structural properties different from those of other codes of French law[15].

Another interesting application of the analysis of networks, particularly relevant for studies on judicial organization, is offered by Katz and Stafford[16] that, based on biographical information about a set of 19,000 law clerks (assistants of judges and federal prosecutors), study the U.S. federal courts as a social system for the purpose, *inter alia*, of providing criteria for measuring the authoritativeness of the careers of the actors of the system and to understand the reasons that generate the differences between the different components of social network.

The list could go on, as in the last few years, the number of publications about SNA and law has grown rapidly[17]. In general terms, we can say that

2005), in "Proceedings of the 12th International Conference on Artificial Intelligence and Law (ICAIL 2009)", available at SSRN: http://ssrn.com/abstract=1419525.

[13] M.J. BOMMARITO II, D.M. KATZ, *A Mathematical Approach to the Study of the United States Code*, 2010, http://arxiv.org/abs/1003.4146; M.J. BOMMARITO II, D.M. KATZ, *Properties of the United States Code Citation Network*, 2009, available at SSRN: http://ssrn.com/abstract=1502927.

[14] R. BOULET, P. MAZZEGA, D. BOURCIER, *Network Analysis of the French Environmental Code*, in Casanovas P., Pagallo U., Sartor G., Ajani G. (eds.), "AI Approaches to the Complexity of Legal Systems", Heidelberg, Springer, 2010, pp. 39-53.

[15] Always with regard to French law, it is worth mentioning the suggestive graphical representation of the network of the relationships between the Articles of the French *Code civil*, offered online by the project Lexmex: http://lexmex.fr.

[16] D.M. KATZ, D.K. STAFFORD, *Hustle and Flow: A Social Network Analysis of the American Federal Judiciary*, in "Ohio State Law Journal", Vol. 71, 2010, n. 3, pp. 457-509.

[17] See, among the others, F.B. CROSS, T.A. SMITH, *The Reagan Revolution in the Network of Law*, available at SSRN: http://ssrn.com/abstract=909217; P.A. HOOK, *Visualizing the Topic Space of the United States Supreme Court*. Indiana Legal Studies Research Paper No. 68, available at SSRN: http://ssrn.com/abstract=948759; U. PAGALLO, *Small World Paradigm and Empirical Research in Legal Ontologies: A Topological Approach*, in Ajani G., Peruginelli G., Sartor G., Tiscornia D. (eds.), "The Multiple Complexity of European Law: Methodologies in Comparison", EPAP, Firenze, 2007, pp. 195-210; R. WINKELS, J. DE RUYTER, H. KROESE, *Determining Authority of Dutch Case Law*, in Atkinson K.M. (ed.), "Legal Knowl-

the spread of network analysis is yielding interesting new insights into the overall structure of the law and into socio-legal phenomena that are relevant in the legal field.

4. Text and Social Network Analysis as Investigative Tools

Among the various application contexts of SNA that are arousing the interest of legal scientists and practitioners, one of the most promising is the study of the social underpinnings of legally relevant phenomena, or, more in detail, the analysis of criminal groups and networks from a structural and functional point of view.

The social dimension plays a crucial role in the evolution of crime: a large part of criminal phenomena from drug networks[18] to international terrorism[19]; from pornography trafficking[20] to hacking and other cybercrimes[21], is strongly conditioned (inhibited or facilitated[22]) by relational dynamics. Criminals are not isolated: they are nested within communities, drawing support from members of the community. In this context, the SNA approach, thanks to the potentialities shown in mapping and measuring the social landscape, is becoming more and more interesting.

edge and Information Systems. JURIX 2011: The 24th Annual Conference", Amsterdam, IOS Press, 2011, pp. 103-112; M. van Opijnen, *Citation Analysis and Beyond: in Search of Indicators Measuring Case Law Importance*, in Schäfer B. (ed.), "Legal Knowledge and Information Systems - JURIX 2012: The 25th Annual Conference", Amsterdam, IOS Press, 2012.

[18] K. Murji, *Markets, Hierarchies and Networks: Race/ethnicity and Drug Distribution*, in "Journal of Drug Issues", Vol. 37, 2007, n. 4, pp. 781-804; Bruinsma, Gerben, Bernasco, Wim, *Criminal Groups and Transnational Illegal Markets: A More Detailed Examination on the Basis of Social Network Theory*, in "Crime Law and Social Change", Vol. 41, 2004, n. 1, pp. 79-94.

[19] V. Krebs, *Mapping Networks of Terrorist Cells*, in "Connections", Vol. 24, 2002, n. 3, pp. 43-52.

[20] J. Johnson, *To Catch a Curious Clicker: A Social Network Analysis of the Online Pornography Industry*, in Boyle K. (ed.), "Everyday Pornographies", London, Routledge, 2010.

[21] D. Décary-Hétu, B. Dupont, *The Social Network of Hackers*, in "Global Crime", 2012, available at SSRN: http://ssrn.com/abstract=2119235.

[22] E. Patacchini, Y. Zenou, *The Strength of Weak Ties in Crime*, in "European Economic Review", Vol. 52, 2008, n. 2, pp. 209-236; D.L. Haynie, *Delinquent Peers Revisited: Does Network Structure Matter?*, in "American Journal of Sociology", Vol. 106, 2001, n. 4, pp. 1013-1057.

Nowadays, the achievement of results by SNA techniques is facilitated by the boundless amount of digital by-products (e-mails, mobile phone calls, credit-card operations, Internet searches, interactions mediated by social networks) generated by social life in the modern world: the digital data-stream and advances in technology are pushing both scholars and law enforcement agencies to figure out new tools and methodologies to illuminate the structures and dynamics of criminal networks.

The idea has recently gained growing popularity: over the past 10 years[23], there has been an increasing number of experiences in the development and in the use of SNA-based investigative software. Criminal Network Analysis[24] (CNA) is a flourishing research area in which criminology, organized crime research, social network theory and computer science converge with other disciplines[25] to offer new insights into crime by means of innovative data mining tools and applications working to unveil hidden patterns in large volumes of crime data and investigative documents.

5. The Case Study

Taking a cue from the scientific scenario so far sketched, we have decided to explore the application of CSS (in particular, information extraction and SNA) techniques in criminal investigation. In this vein, fundamental in-

[23] X. Shang, Y. Yuan, *Social Network Analysis in Multiple Social Networks Data for Criminal Group Discovery*, in "International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery", 2012, pp. 27-30; J.A. Johnson, J.D. Reitzel, *Social Network Analysis in an Operational Environment: Defining the Utility of a Network Approach for Crime Analysis Using the Richmond City Police Department as a Case Study*, IPES, Coginta and DCAF Working Paper No. 39, 2011; N. Memon, H.L. Larsen, *Investigative Data Mining Toolkit: A Software Prototype for Visualizing, Analyzing and Destabilizing Terrorist Networks*, in "Visualising Network Information", http://www.dtic.mil/dtic/tr/fulltext/u2/a477075.pdf; Y. Lu, M. Polgar, Y. Cao, *Social Network Analysis of a Criminal Hacker Community*, in "Journal of Computer Information Systems", Vol. 51, 2010, n. 2, pp. 31-41; J. Xu, H. Chen, *CrimeNet Explorer: A Framework for Criminal Network Knowledge Discovery*, in "ACM Transactions on Information Systems", Vol. 23, 2005, n. 2, pp. 201-226; J. Xu, H. Chen, *Criminal Network Analysis and Visualization*, in "Communications of the ACM", Vol. 48, 2005, n. 6, pp. 100-107.

[24] For a compelling introduction to the research area of criminal network analysis, see C. Morselli, *Inside Criminal Networks. Studies of Organized Crime*, Berlin-Heidelberg, Springer, 2010.

[25] K. von Lampe, *The Interdisciplinary Dimensions of the Study of Organized Crime*, in "Trends in Organized Crime", Vol. 9, 2006, n. 3, pp. 77-95.

puts came from the collaboration with Italian deputy prosecutors that highlighted two circumstances relevant for our research:

- If we put aside more "traditional" information systems (databases containing complaints, criminal records, police reports, etc.), criminal investigation remains primarily a manual process. Available computational tools do not provide either advanced information extraction functionalities or structural analysis of network knowledge from criminal justice documents. Furthermore, investigative information, even when in digital format, is not structured.
- Tools allowing us to gather, analyze, visualize and extract information from investigative documents produced by the prosecutor's office and police would be considered extremely useful. Particular attention has been aroused by visual and "scientifically grounded" tools illuminating not only the extension, the topology and the patterns of groups of people involved in illegal activities, but also the role of specific individuals within the organization. After all, knowledge about the structural and functional properties of criminal networks is fundamental for both investigation and the development of effective prevention strategies.

Starting from the consideration of these circumstances, we have initiated a project in which domain experts (lawyers and judges) collaborate with computer scientists to develop and test an integrated environment based on social network analysis techniques able to:

- support the automatic retrieval and marking up of information necessary for performing network analysis of criminality;
- generate and analyze graphs of criminal networks under investigation;
- provide a diachronic view of the evolution of criminal networks;
- organize and manage visually the investigative material;
- make predictions about the potential of an individual to belong to a criminal group.

In the light of the guidance given by the prosecutors involved in the project, an environment of this kind not only could offer insights into criminal phenomena difficult to obtain in any other way, but could also have an impact in the dynamic of the investigation and trial. The graphs and the results of analysis carried out on them could be a tool to facilitate information sharing and collaboration between the judges and the investigating police; moreover, they could be used to support requests for commital for trial with the evidence of a scientific kind with particular value for the argumentation.

In the part of the project carried out so far, after having examined the characteristics of the documents produced in the course of the investigation by prosecutors, we have focused our attention on the extraction of information from a particular type of pleading (request for an arrest warrant) and the consequent application of SNA metrics to this information.

The starting point of the project was offered by an interesting research[26] by a sociologist of deviance who not only brought us closer to the use of SNA in the study of criminal phenomena but also aided us in the retrieval of documents on which to start experimenting. Oriented towards purely scientific purposes, the work of Scaglione was to analyze the characteristics of the social network emerging from the wire-tapping reported in two requests for precautionary measures[27] to be taken that had been subjected to the manual tagging of parts of the texts deemed to be relevant.

Using one of the procedural documents examined by Scaglione, we started our case study, by posing as its first objective the replication of his findings in the study of the criminal network and, at the same time, beginning to implement the various components of the tool supporting investigations imagined.

In our case study, whose properties are listed in Tab. 1, the analyzed document contains two kinds of information source:

- *Telephone tapping*: this is a transcript of a phone call started or reaching a person under investigation.
- *Wiretap*: a transcription of dialogues in an environment, there is not a person who starts the call. In this case, we assume that each person that is shown in the conversation is communicating with any other person involved in wiretap.

5.1. Implementation

Before starting with details about the implementation, we have to emphasize that an obstacle in this area is the difficulty of finding scanned docu-

[26] A. SCAGLIONE, *Reti mafiose. Cosa Nostra e Camorra: organizzazioni criminali a confronto*, Milano, Franco Angeli, 2011.

[27] Data used for the construction of the graph by Scaglione were manually extracted from unstructured documents. No consideration has been given to the content of the calls: the data taken into account for the construction of the graphs have only been the following: calls (two nodes are connected if there is no communication between them); the number of contacts; the direction of phone call (who calls who).

Characteristics	Value
Document type	Remand document
Organization name	Clan Cava
Organization class	Camorra
People involved	73
Location	Quindici, Avellino, Italy
Pages of the document	ca 3000
Number of telephone and wiretap	2791
Number of phone under tapping	300

Tab. 1 – Properties of the case study

ments, or other documents that are readily convertible into formats that can then be adapted to automatic analysis.

As anticipated, we have analyzed the remand document used by Scaglione for his work on the comparison of criminal organizations. This document contains information about nearly 150 criminals classified in two *Families*: the *Cava Family* from Quindici (Avellino) and the *Rinzivillo Family* from Gela (Caltanissetta).

To represent the crime network described by Scaglione we have used a graph, one of the already well-known ways to represent and investigate criminal network[28]. A graph is composed of a pair of sets, called nodes and edges, where edges link the nodes together. Nodes represent people, groups or organizations (but also vehicles, building and so on), that are connected through social ties (that is, the edges) in which a variety of resources are exchanged or used. In order to build our crime network we have taken into account the following information:

- telephone tapping and wiretap (nodes are connected if there was a communication between them)
- the phone call direction (who calls whom)

The phone call direction could be interesting for deriving the importance of members of the organizations. We have to emphasize that we did not consider the content of the call, as our main and initial interest was in the struc-

[28] J. XU, H. CHEN, *Criminal Network Analysis and Visualization*, cit.; J. SCHROEDER, J. XU, H. CHEN, *CrimeLink Explorer: Using Domain Knowledge to Facilitate Automated Crime Association Analysis*, in "Proceedings of the 1st NSF/NIJ conference on Intelligence and security informatics", 2003, pp. 168-180.

ture of the resultant network. We have organized the work of information extraction and analysis in three different phases, as shown in Fig. 1.

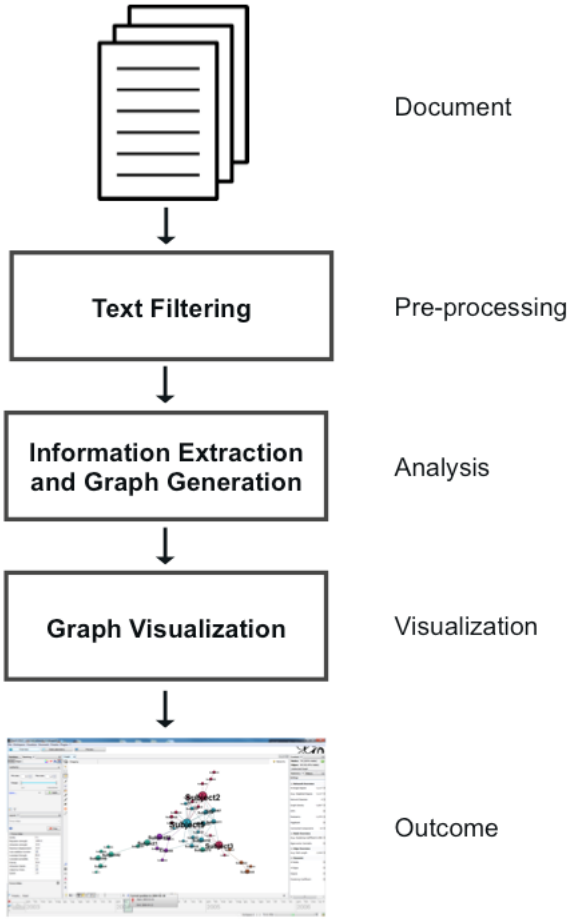

Fig. 1 – Case study workflow

Firstly, we needed to apply an initial "*Text Filtering*" phase to make official documents suitable for the subsequent automatic "*Information Extraction and Graph Generation*" phase. In this second phase, the remand documents were automatically processed to analyze behaviors and relationships

in order to produce, in the last phase (namely, *Graph Visualization*), the corresponding visual representation.

The way we analyzed and visualized the network can be classified according to a taxonomy defined by Klerks[29] who has analyzed and evaluated the approach for criminal network analysis, classifying them in:

- *First generation*
  A completely manual task. It uses a matrix where each row and column represents a person (an offender) and the value is the number of established contacts. Eventually an image can be drawn from the matrix.
- *Second generation*
  Can automatically visualize a representation of the criminal network, adjusting the position of nodes to achieve a better visualization.
- *Third generation*
  An evolution of the second that moves the focus onto the social context. Focusing on the social context could help the investigators in their search for high ranked members of the organization, to find out how much power members have within the organization, important intermediate or even weak points in the organization structure. Furthermore, it could help to clarify how recruitment works and how the order and/or knowledge are transferred by one member to another.

The third generation approach makes it easier to give answers to qualitative questions, for example, how much the member of the organization communicates before and after committing a crime, and what are the hot topics in their conversations. Our approach can be classified as a third generation approach.

5.2. Text Filtering

In this phase, we filtered out unnecessary information and marked some specific parts needful for the network analysis. To this end, we have developed, in Perl[30] a simple parser that extracts entities (name, surname, telephone number, etc.) relevant for the graph. It was not possible to completely

[29] P. KLERKS, *The Network Paradigm Applied to Criminal Organisations: Theoretical Nitpicking or a Relevant Doctrine for Investigators? Recent Developments in the Netherlands*, in "Connections", Vol. 24, 2001, n. 3, pp. 53-65.

[30] Perl is a high-level, general-purpose programming language providing powerful text processing facilities.

process the documents automatically because of their nature. In fact, they are known to suffer from the following critical deficiencies[31]:

- *Incompleteness*: sometimes some information about nodes and relations are missing, making ii impossible to construct a the global vision of the network that resembles as much as possible the real one. Examples include the minimization of the criminals' interactions to avoid attracting police attention and the hiding of their interactions behind various illicit activities.
- *Incorrectness*: Incorrect data regarding criminals' identities, physical characteristics, and addresses may result either from unintentional data entry errors or from intentional deception by criminals. These types of errors can be easily addressed.
- *Inconsistency*: Information about a criminal who has multiple police contacts may be entered into law enforcement databases multiple times. These records are not necessarily consistent, involving inaccuracies when building the network.

5.3. Information Extraction and Graph Generation

In this phase, all entities have to be extracted and used to find out relations between the members of the organization. To do that, our software firstly extracts from the remand document (in a specific section) the list of people involved in the judiciary investigation. They will represent the nodes of our network. Next, the rest of the document is parsed and for each wiretap transcription:

1. We check if the people involved match against the previously defined list
2. If a match exists, we add a new edge between the corresponding entities in the graph.

Our software produces an XML file representing the network and complying with the GEXF file format specifications[32], a language for describing complex networks structures, their associated data and dynamics.

[31] J. XU, H. CHEN, *Criminal Network Analysis and Visualization*, cit.

[32] See http://gexf.net/format.

5.4. Graph Visualization

This last phase allowed us to visualize the resultant graph. We loaded the produced GEXF file into a framework for graph analysis and visualization named *Gephi*[33], a framework that incorporates and offers a wide range of algorithms from Graph Theory literature, including: algorithms for extracting metrics (statistical properties); algorithms for group discovery; algorithms for visualization.

Once loaded our graph, we also computed some statistics, such as, for example, the importance of the organizations' members. The graph representing our case study is shown in Fig. 2. Nodes represent members (we used Subject IDs instead of real full names) and their color and size give information about group membership and authoritativeness, respectively.

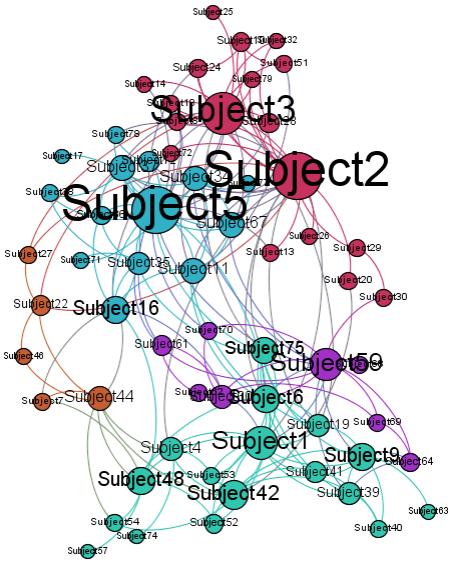

Fig. 2 – Criminal network graph

[33] M. BASTIAN, S. HEYMANN, M. JACOMY, *Gephi: An Open Source Software for Exploring and Manipulating Networks*, in "International AAAI Conference on Weblogs and Social Media (2009)". Other tools used in social network analysis include Ucinet, Pajek, NetMiner, Stocnet, NodeXL. For a general overview about network analysis tools see A. TROBIA, V. MILIA, *Social Network Analysis. Approcci, Tecniche, Applicazioni*, Roma, Carocci, 2011, p. 165.

Another interesting feature of the tool so far developed is the possibility to switch from a static to a dynamic version of the graph (see Fig. 3 and Fig. 4). Taking into account the date associated to every wiretap, the software is able to generate different overviews of the organization under investigation allowing researchers to examine the evolution of the criminal group over time, identifying temporal patterns and trends (growth, decline of criminal groups).

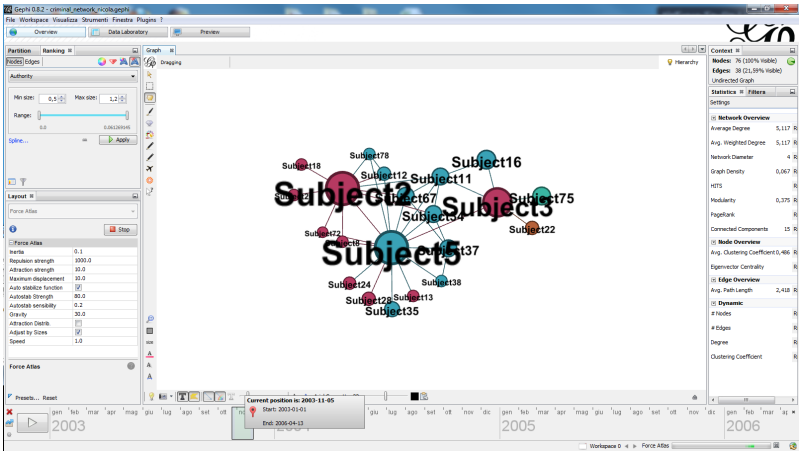

Fig. 3 – Screenshot of Gephi interface with timeline slider

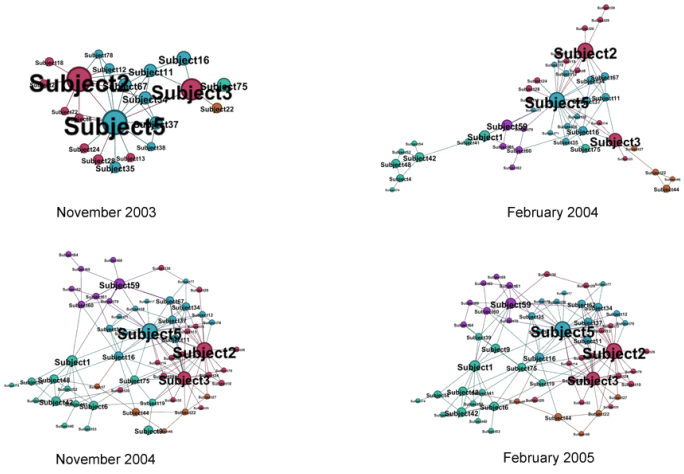

Fig. 4 – Evolution of a criminal network over time

6. Conclusions

Although in progress, experiments have given some preliminary outcomes. The most tangible results are found on the level of implementation. The work done so far has allowed us to explore the issues related to the parsing of unstructured investigative and prosecution documents and to obtain useful information on the processing of these kinds of documents. The analysis carried out are encouraging in terms of reliability in extracting relevant information: data derived from automatic analysis and the subsequent processing of the documents fit with the trial evidence and have generated graphs whose measures are essentially comparable to those resulting from the work of Scaglione that was taken as a reference. To all this, always on the application level, there is also the implementation of timeline for the diachronic view of the criminal network. From the point of view of the analysis of the phenomenon being investigated, the research suffers from a number of limitations arising primarily from the characteristics of the documentary material used: a first limit is represented by the fact of having used a single document even if meaty, another aspect – that is not secondary – is the nature of the information processed. The number and the direction of phone calls used for the construction of the graph, by itself, is not sufficiently and univocally meaningful: depending on the group, the region and the culture considered, a person playing a secondary role in a criminal organization can be the recipient of a large number of phone calls. The issue that came out during the discussion with the prosecutors emphasized how that which turns out to be relevant on the scientific level of a sociological-type study, albeit inspiring, often may not be significant for investigation purposes. According to this consideration, it clearly emerged how the creation of a support tool for carrying out an investigation requires not only increasing the amount and variety of information processed to build up graphs (e.g., gathering data from criminal justice databases) but also refining criteria used in processing judicial and police information trying to "embed" the inquirer's know-how in information extraction procedures. In this perspective, it will be useful to use reliable methods to weigh the impact that investigative information gathered have on the probability that an individual is part of a group or a criminal organization.

The most interesting result, however, is found on the methodological level; the interdisciplinary collaboration regarding crime analysis technologies is showing that it can have an impact on the development of interdiction and law enforcement strategies.

The research will continue in two directions devoting itself, on the one hand, to a deepening of the scientific issues linked to social and criminal network analysis and, on the other, to the design and implementation of new tools for the collection and analysis of investigative data. A crucial first step in this direction seems to be to design a tool that allows us to collect and mark up *ab initio* with the necessary metadata information to be used for building graphs.

The road ahead seems to be challenging both, on the one hand, for Criminology and Criminal law, and, on the other hand, for Legal informatics. For Criminology and Criminal Law, information technologies and SNA techniques represent an opportunity for methodological enrichment through an accentuation of the quantitative dimension of the study of crime. For Legal informatics, the topic not only represents a new application context but also the opportunity to build on and improve the knowledge already produced in processing legal information and documentation. In the future, moreover, there is the possibility of integration with other CSS methods, for example, with social simulations that may serve to combine SNA with the generative perspective of simulations to which the law has also begun to approach[34]. The study of crime and, eventually, the prediction of its evolution can be based not only on the techniques of inferential statistics applied to large amounts of data[35] but also on methods that take better account of the cognitive, social and institutional processes underlying criminal phenomena[36].

[34] See, for example, the special Issue dedicated to *Simulation, Norms and Laws*, of "Artificial Intelligence and Law", Vol. 20, 2012, n. 4 and Vol. 21, 2013, n. 1.

[35] A particularly suggestive example of this technique is provided by *PredPol*, a software environment for "predictive policing" developed under the guidance of anthropologist P.J. Brantingham of UCLA (http://www.predpol.com/).

[36] As numerous and recent publications demonstrate, not only is the intersection between social simulations and network analysis a reality (for an interesting example regarding market dynamics see K. LEE, S. KIM, C.O. KIM, T. PARK, *An Agent-Based Competitive Product Diffusion Model for the Estimation and Sensitivity Analysis of Social Network Structure and Purchase Time Distribution*, in "Journal of Artificial Societies and Social Simulation", Vol. 16, 2013, n. 1, http://jasss.soc.surrey.ac.uk/16/1/3.html and the bibliography cited in it) but also the simulation study of crimes for the purpose of prediction (see, i.a., M. FONOBEROVA, V.A. FONOBEROV, I. MEZIC, J. MEZIC, P.J. BRANTINGHAM, *Nonlinear Dynamics of Crime and Violence in Urban Settings*, in "Journal of Artificial Societies and Social Simulation", Vol. 15, 2012, n. 1, http://jasss.soc.surrey.ac.uk/15/1/2.html; N. MALLESON, P. BRANTINGHAM, *Prototype Burglary Simulations for Crime Reduction and Forecasting*, in "Crime Patterns and Analysis", Vol. 2, 2009, n. 1, pp. 47-66).